DOCUMENT RESUME

ED 246 091 TM 840 364

AUTHOR Jaeger, Richard M.; Busch, John Christian
TITLE The Effects of a Delphi Modification of the Angoff-Jaeger Standard-Setting Procedure on Standards Recommended for the National Teacher Examinations.
PUB DATE Apr 84
NOTE 27p.; Paper presented at the Joint Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, (New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Standards; Admission Criteria; *College Entrance Examinations; *Cutting Scores; Decision Making; *Delphi Technique; *Discussion; Higher Education; Teacher Education; Testing
IDENTIFIERS Angoff Jaeger Method; *National Teacher Examinations; North Carolina; *Standard Setting

ABSTRACT

A standard-setting experiment was conducted in which one group of participants was instructed to refrain from discussing their recommended test standards and another group was requested to report the reasons underlying their recommendations. The procedures used were designed to induce convergence of recommendations, but did not require consensus. A structured judgment procedure was used to obtain recommendations on appropriate standards of performance on the Reading and Social Studies subtests of the National Teacher Examinations. Recommended standards of performance were provided for applicants to North Carolina's teacher education programs, by public school teachers and college and university faculty. Each participant was asked to complete the subtest, then to estimate, for each subtest item, the probability that a minimally competent examinee would be able to answer the item correctly. Using the Delphi Technique, participants received feedback about the ratings and then were asked to reconsider their initial recommendations. One group was asked not to discuss the feedback; in the other, discussion was encouraged. The reduction in variance from the first judgment session to the second was substantially greater for the discussion group. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED246091

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

THE EFFECTS OF A DELPHI MODIFICATION OF THE ANGOFF-JAEGER STANDARD-SETTING PROCEDURE ON STANDARDS RECOMMENDED FOR THE NATIONAL TEACHER EXAMINATIONS

Richard M. Jaeger and John Christian Busch
University of North Carolina at Greensboro

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. M. Jaeger
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 840 369

Presented at the 1984 Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, April 23-27, 1984, in a symposium entitled: "Setting Standards for Teacher Certification: Recent Advances in Methodology."

BACKGROUND AND RATIONALE

Virtually all standard-setting methods based on judgment of test items described in recent reviews (Glass, 1978; Hambleton, 1980; Hambleton and Eignor, 1980; Hambleton, et al., 1978; Jaeger, 1976; 1979; Livingston and Zieky, 1982; Meskauskas, 1976; Shepard, 1980) involve what social psychologists term "coworking groups" (Allport, 1924; 1962). That is, individuals making judgments on appropriate test standards typically work in one another's presence, but without explicit communication on the issue to be judged. The work of Allport and others suggests that coworking groups produce more consistent judgments than do individuals working alone, even when members of the group do not communicate.

A large body of social psychological research, beginning with the work of Sherif (1947), and including that of Festinger (1954) and Schachter (1959), suggests that having individuals share their judgments aloud will produce a shift in judgments toward convergence. Balance theory (Abelson and Rosenberg, 1958; Cartwright and Harary, 1956; Heider, 1944; 1946; 1958;), as well as the influence of normative expectations (Myers, et al., 1974; Wallach and Kogan, 1965), have been proposed as explanations of this phenomenon.

Although some standard-setting research has required participants to state the reasons for their judgments (Koffler, 1980) or discuss their judgments to the point of consensus (Brennan and Lockwood, 1980), theeffects of such procedures have not been studied systematically.

This paper reports the results of a standard-setting experiment in which one group of participants was instructed to refrain from discussing their recommended test standards and another group was requested to report the reasons underlying their recommendations. The procedures used were designed to induce convergence of recommendations, but did not require consensus.

When consensus is required, it has often been found that dominant individuals will unduly influence the judgments of the group and that the decision of the group will be more extreme than the average position of the individuals that compose it (Marquis, 1962; Stoner, 1961; Wallach and Kogan, 1959). Indeed, Brennan and Lockwood (1980) found evidence of dominance by two of five participants in their standard-setting research and a shift toward a higher recomended standard when they evaluated the Nedelsky (1954) standard-setting procedure.

The principal advantage of a standard-setting procedure that induces convergence, without necessarily affecting the mean recommended standard, is more-precise estimation of the mean. This statement presumes what Brennan and Lockwood (1980) termed a random-effects model in which participants in the standard-setting process are regarded as a sample, selected randomly from a potential population. It is the population mean recommended standard that is to be estimated.

Beyond its potential for statistical efficiency, a standard-setting procedure that invites participants to describe the rationale underlying their judgments has an added advantage. It is consistent with the view that standard-setters should be informed decision-makers who are fully cognizant of the implica-

tions of their recommendations. It is reasonable to presume that, when test items are reviewed, some members of a group will have insights that do not occur to others. Discussion of these insights cannot help but result in better-informed decision making.

METHODS

Overview

A structured judgment procedure was used to obtain recommendations on appropriate standards of performance on the Reading and Social Studies subtests of the National Teacher Examinations. Recommended standards of performance were provided for applicants to North Carolina's teacher education programs, by public school teachers and college and university faculty.

Each participant in the study was first asked to complete the subtest that was the subject of his/her recommendations. Participants were then given two separate opportunities to provide recommendations on appropriate standards. In the first, they followed procedures attributed to Angoff (1971). They reviewed each test item, and then estimated the probability that a minimally competent applicant to a teacher education program could answer the item correctly. In the second, they were given a variety of data on examinees' actual test performances, following procedures suggested by Jaeger (1978, 1982), and then reconsidered their original item-by-item recommendations.

Participants reviewing each subtest were assigned randomly to one of two groups. One of these groups (termed the "Silence Group") was instructed to refrain from discussing their recom-

mendations prior to, or during, the second judgment session. The other group (termed the "Discussion Group") was instructed to discuss their initial recommendations, under the strict guidance of a group leader, prior to reconsidering those recommendations.

A recommended standard was computed for each participant following each judgment session. These standards were summarized using appropriate descriptive statistics, and were then used in several inferential analyses, to examine how, if at all, recommended standards varied across the Silence and Discussion procedures, judgment sessions, and types of participants.

A more-detailed discussion of these methods follows.

Sampling and Assignment of Participants

Participants in this study consisted of 14 college and university faculty members and 14 public school teachers. College and university faculty were randomly sampled from selected departments in five public and private colleges and universities in and near Greensboro, North Carolina. Faculty assigned to the Reading subtest were sampled from departments of English, college-level remedial reading programs, or English education and reading education programs in departments of education. Faculty assigned to the Social Studies subtest were sampled from departments of history, sociology, anthropology, psychology, or education. Education faculty were restricted to those who taught courses in social foundations, social studies education, or educational psychology.

Public school teachers who participated in the study were sampled from the three school systems in Guilford County, North Carolina. To qualify for selection, teachers had to hold a cur-

rent certificate, be employed full time, be a graduate of a North Carolina teacher education program, and have between two and five years of teaching experience.

Six public school teachers were randomly assigned to the Silence and Discussion Groups that were to review the Reading subtest, and the remaining eight public school teachers were randomly assigned to the Silence and Discussion Groups that were to review the Social Studies subtest.

Eight college and university faculty with backgrounds in reading or English were randomly divided between the Silence and Discussion Groups that were to review the Reading subtest. The remaining six college and university faculty members were randomly divided between the Silence and Discussion Groups that were to review the Social Studies subtest. To provide a larger group for discussion, four of these faculty members were assigned to the Discussion Group and only two were assigned to the Silence group.

Instruments

The Communication Skills Test of the National Teacher Examinations contains four subtests. Three multiple-choice subtests assess examinees' skills in reading, writing mechanics, and listening. The fourth subtest assesses skills in writing production. Only the Reading subtest, a 30-minute, 30-item multiple-choice test, was used in this study.

The General Knowledge Test of the National Teacher Examinations contains four multiple-choice subtests that assess examinees' knowledge in mathematics, social studies, science, and

literature and fine arts. Only the 30-minute, 30-item Social Studies subtest was used in this study.

These subtests were designed by Educational Testing Service to be used in assessing the skills and knowledge of beginning teachers or applicants for teacher certification. The North Carolina State Board of Education has adopted the General Knowledge and Communication Skills Tests for use in screening applicants to the state's teacher education programs. The study described here was part of a larger investigation of test validity and appropriate test standards. Data collected in that investigation provided internal consistency estimates of subtest reliability in the low-to-mid 80's.

Procedures

Participants in each group were first given 30 minutes to complete the subtest for which they recommended standards of performance. The purpose of this activity was to familiarize each participant with the content of the subtests, from the perspective of an examinee. After they had completed their subtest, each participant was asked to estimate, for each subtest item, the probability that a minimally competent examinee would be able to answer the item correctly. More specifically, participants were told to do the following:

> "You are to imagine a hypothetical teacher education applicant who has the minimum amount of knowledge and skill in the subject area covered by the [Reading, Social Studies] section of the [Communication Skills, General Knowledge] Test, necessary to succeed in the approved North Carolina teacher education program with which you are most familiar. This hypothetical student has minimally acceptable, but not outstanding, skills and knowledge and might be what we call a low 'C' student. You are to estimate, for each item in the test section, the probability or 'chances' that this mini-

mally competent applicant would be able to answer the test item correctly."

Participants were then given instruction on the meaning of probability, and were told to record each of their judgments on a response sheet by marking one of ten circles, numbered from one to ten. They were told that the circle numbered one represented probabilities in the range zero to 0.10; that the circle numbered two represented probabilities in the range 0.11 to 0.20; and so on, up to the circle numbered ten, that represented probabilities in the range 0.91 to one.

In accordance with Angoff's (1971) procedure, the probabilities estimated for each item by a participant were summed to yield a recommended standard for that participant.

In a second session, all participants received instruction on the meaning and interpretation of item difficulty values (p-values) and cumulative distribution functions. They then used simulated item difficulty data and a simulated cumulative distribution of examinees' performances on a test composed of ten sample items in reading and social studies, to provide item-by-item estimates of the probabilities of success of teacher education applicants. The purpose of this exercise was to provide instruction and experience in using data when considering the probability that a teacher education applicant would answer a test item correctly.

Following the instruction session, all participants were given the opportunity to reconsider their initial estimates of the probability that a minimally competent applicant would answer each subtest item correctly. As a judgmental aid, participants

were provided with four kinds of data: (1) an estimated difficulty value (p-value) for each item, based on previous administrations of the item by Educational Testing Service, (2) an estimated difficulty value (p-value) for each item, applicable to a subpopulation of examinees with total scores that were close to the passing scores established in previous studies of the National Teacher Examinations, (3) the cumulative distribution function of subtest scores, based on previous administrations of the subtest by Educational Testing Service, and (4) their response sheet from the first judgment session, with the recommended standard derived from their item-by-item judgments clearly indicated. Participants were instructed to consider these data as though they were representative of the performance of applicants to North Carolina's teacher education programs.

At this point, participants assigned to the "Silence" groups were asked to reconsider their initial recommendations and then complete a new response form by circling one of the numbers from one to ten for each item. They were reminded of the nature of the desired judgments and the meaning of each scale value in terms of probabilities. These participants were requested to refrain from any discussion of their recommendations until their response sheets had been collected by their group leader.

Participants assigned to the "Discussion" groups were given the same data as those assigned to the "Silence" groups. They were then asked to participate in a controlled discussion of the reasoning underlying each of their first-session judgments.

Both the Reading subtest and the Social Studies subtest are composed of items that can be grouped because of their content or

because they share a common stimulus. Leaders of the "Discussion" groups used these item groupings to organize discussion of sets of test items. Participants in these groups were told the following:

> "It is important to realize that there will be some diversity in the probability ratings that different judges render and that individuals have different reasons for their judgments. First we will ask you to review the ratings you gave to items [1 and 2 or 1,2,3] during the first session. We will ask you to discuss your own choices and your reasons for them. Finally, we'll ask you to reconsider your first judgments in light of the statistical information you've received and your discussion with your colleagues."
>
> "... why don't we locate items [1 and 2 or 1,2,3] on the test booklet and the judgments you made during the first standard-setting session. These can be found on your judgment form from the first session. Let's take a moment to compare our judgments. Is there anyone who would care to share their judgment with the group? Did anyone set their probabilities 'high,' say greater than 6 on the 1 to 10 scale for items [1 and 2, or 1,2,3]? [Leader should ask the respondent to describe why that probability level was used]."
>
> "Probably other people made different decisions. Did anyone set their probabilities 'low,' say less than 4 on the 1 to 10 scale?" [Leader should ask the respondent to describe why that probability level was used]."
>
> "Given your colleagues' opinions, the statistics for each item, and the information about the percentage of applicants who would fail if a particular standard was set, next please mark your estimate of the probability that a hypothetical minimally competent applicant would answer each of these items correctly. Use the response form just as you did in the first session."

The same procedure was then used with each set of items. Particpants in the "Discussion" groups openly discussed the reasons they had estimated both high and low probabilities for each set of items, and then provided new (and possibly different) estimates for those items.

A recommended standard was computed for each participant

just as in the first judgment session, by summing the midpoint's of the intervals corresponding to that participant's choices on the ten-point probability scale.

Analysis

Several descriptive and inferential statistical procedures were used to compare the central tendency and variability of standards recommended by the following groups: (1) participants in the "Silence" groups and participants in the "Discussion" groups, (2) college and university faculty and public school teachers, and (3) recommendations provided during the first judgment session and recommendations provided during the second judgment session. In addition to computing means and variances of recommended standards for each of these groups and conditions, we completed a three-way repeated measures analysis of variance using the standards recommended for each subtest. We also used Bartlett's test to investigate the significance of changes in the variability of recommended standards.

RESULTS

Descriptive Analyses

Means and variances of standards recommended by each group are shown in Tables 1 and 2 for the Reading subtest and the Social Studies subtest, respectively. When rounded to integers, mean standards recommended for the Reading subtest by each group were quite similar; ranging from 13 based on data from the first judgment session to 16 based on data from the second. With the exception of college and university faculty assigned to the "Discussion" group for the Social Studies subtest, mean recom-

mended standards for the Social Studies subtest were similar to those recommended for the Reading subtest; they ranged from 13 to 17.

INSERT TABLES 1 AND 2 ABOUT HERE

Recommended standards based on second-session judgments were consistently less variable than those based on first-session judgments, for all groups of participants and both subtests. In most cases, the reduction in variance from the first to the second judgment session was substantial. However, as we shall note later, the reduction in variance for the "Discussion" groups was noticeably greater than that for corresponding "Silence" groups.

Although the iterative judgment procedure resulted in reduced variability among recommended standards, it did not appear to have a significant effect on the mean recommended standards. As shown in Table 1, the mean standard recomended by all participants who reviewed the Reading subtest increased by only four-hundredths of a point from the first judgment session to the second. The corresponding change in the mean standard recomended for the Social Studies subtest was just over half a point (Table 2). Among the eight groups of participants who provided recommendations on these two subtests, five recommended a lower mean standard during the second judgment session and three recommended a higher mean standard during the second judgment session; consistent shifts in mean recommended standards were not apparent.

The reader should keep in mind the small sample sizes repre-

sented by the eight groups for which data are reported in Tables 1 and 2. Because of these sample sizes, estimates of means and variances for individual groups are imprecise. Only the consistency of results across groups is compelling and worthy of interpretation.

Inferential Analyses

Separate three-way repeated measures analyses of variance were conducted using standards recommended for the Reading subtest and the Social Studies subtest. The results of these analyses are shown in Table 3. Type of treatment ("Discussion" and "Silence") and type of participant (public school teacher and college and university faculty) were used as between-subjects factors. Session number was used as a within-subjects factor in each analysis.

INSERT TABLE 3 ABOUT HERE

The results of these analyses can be summarized as follows:

(1) For both the Reading subtest and the Social Studies subtest, the mean standard recommended by the "Discussion" group did not differ significantly from the mean standard recommended by the "Silence" group ($p > .10$).

(2) For both the Reading subtest and the Social Studies subtest, the mean standard recommended by public school teachers did not differ significantly from the mean standard recommended by college and university faculty ($p > .10$).

(3) For both the Reading subtest and the Social Studies subtest, the mean standard recommended during the first judgment

session did not differ significantly from the mean standard recommended during the second judgment session ($p > .10$).

(4) For the Social Studies subtest, there was a significant interaction ($p < .05$) between type of participant and type of treatment. This interaction is discussed below.

(5) For the Social Studies subtest, there was a significant three-way interaction ($p < .10$) between type of participant, type of treatment, and session number.

(6) For the Reading subtest, none of the interaction terms in the three-way analysis were statistically significant.

Interactions are often best understood when means obtained under a variety of conditions are shown on the same graph. We provide such a graph in Figure 1, for the significant interactions of standards recommended for the Social Studies subtest. From this figure it is clear that the college and university faculty who participated in the "Discussion" group recommended standards that were substantially lower than those recommended by faculty who participated in the "Silence" group. However, both of these gvoups recommended a higher mean standard during the second judgment session than during the first. Public school teachers who participated in the "Discussion" group recommended mean standards that were similar to those recommended by teachers who participated in the "Silence" group. However, the mean recommendations of public school teachers who participated in the "Discussion" group were considerably higher than those of college and university faculty who participated in the "Discussion" group, whereas this pattern was reversed for participants in the "Silence" group. Finally, although the mean standard recommended

by public school teachers who participated in the "Discussion" group was lower during the second session than during the first, the mean standard recommended by public school teachers who participated in the "Silence" group was higher during the second session than during the first. Once again, we must warn the reader that these results are based on very small group sizes and, although statistically significant interaction effects were found, changes in the judgments of two or three participants could affect these findings materially.

INSERT FIGURE 1 ABOUT HERE

Although the effects of various factors on mean recommended standards was of more than passing interest in this study, we were more interested in the effects of these factors on the variability of recommended standards. If the variability of recommended standards can be reduced, the mean recomended standard will be estimated with greater precision. Procedures that can effect such a reduction might lead to substantial economies, since smaller samples of participants would be needed to estimate mean recommended standards at a specified level of precision.

Since the power of conventional statistical tests to detect differences in population variances is quite low for samples of the sizes assigned to individual groups in this study, we are not distressed by finding differences in sample variances of standards recommended by the Discussion group and the Silence Group, to be non-significant ($p > 0.10$), for each subtest and each data-collection session. Null hypotheses on the homoscedasticity of standards recommended by public school teachers and college and

university faculty were also retained ($p > 0.10$).

We noted above that the variability of standards recommended as a result of the second judgment session was lower than the variability of standards recommended as a result of the first judgment session, for both the Reading subtest and the Social Studies subtest. When standards recommended by participants who reviewed the Reading subtest were pooled across the Discussion and Silence groups, the first-session variance was 9.67 and the second-session variance was 2.91. Based on a sample size of 14 and a reduction in sample variance by a factor of 3.3, a null hypothesis of homoscedasticity across the two judgment sessions would be rejected ($p < 0.2$). Results for the Social Studies subtest are similar. The first-session variance was 24.34 and the second-session variance was 11.42. The reduction in variance by a factor of 2.1 was statistically significant ($p < 0.05$), based on a sample size of 14.

The results of greatest interest here were the comparative reductions in the variance of recommended standards brought about by the Discussion and Silence treatments. For both the Reading subtest and the Social Studies subtest, the reduction in variance from the first judgment session to the second was substantially greater for the Discussion group than for the Silence group. For the Discussion group reviewing the Reading subtest, the first-session variance of recommended standards was 14.85 and the second-session variance was 3.94. Variance was reduced by a factor of 3.87, from the first session to the second. For the Silence group reviewing the Reading subtest, the first-session

variance was 5.13 and the second-session variance was 2.23. Variance was reduced by a factor of 2.30 from the first session to the second. It is evident that the reduction in variability of recommendations, from the first session to the second, was far greater for the Discussion group than for the Silence group. In fact, the ratio of variance ratios, 3.87 divided by 2.30, equals 1.68.

Results obtained for the Social Studies subtest are consistent with those found for the Reading subtest. The variance of standards recommended by members of the Discussion group during the first judgment session was 28.16, and the variance of standards recommended during the second judgment session was only 6.60. Thus variability was reduced by a factor of 4.27 from the first judgment session to the second. The variance of standards recommended by members of the Silence group was 21.32 during the first judgment session and 10.27 during the second judgment session. For this group, variance was also reduced from the first judgment session to the second, but the ratio is only 2.08. Again, computing the ratio of variance ratios reveals that the reduction in variance from the first judgment session to the second was 2.05 times greater for the Discussion group than for the Silence group.

Although we have not investigated the statistical significance of these comparative reductions in variances (the small sample sizes involved would provide extremely low power for testing the equality of variance ratios), the consistency of these results across subtests encourages a belief that the Discussion treatment resulted in greater movement toward homogeneity

of judgments than did the Silence treatment.

CONCLUSIONS

The results described above supported a decision to use the "Discussion" treatment in a statewide standard-setting study involving 306 participants from public schools, colleges, and universities throughout North Carolina. As a part of that study (to be reported elsewhere) we asked participants whether they felt that their recommendations had been influenced by the discussions they heard or contributed to. Fifty-nine percent of the participants agreed, or strongly agreed, that the discussions had influenced their judgments. This value is only slightly lower than the 63 percent who reported that their judgments had been influenced by knowing the percentages of examinees who answered each test item correctly.

The results of the statewide study also show that a combination of normative data and discussion, coupled with an opportunity to reconsider initial recommendations, produces a marked convergence in particpants' recommended test standards.

In total, this evidence suggests that a discussion procedure of the sort examined in this study is a useful element of an item-based standard-setting procedure.

REFERENCES

Abelson, R. P. & Rosenberg, M. J. Symbolic psychologic: A model of attitudinal cognition. Behavioral Science, 1958, 3, 1-13.

Allport, F. H. Social Psychology, Boston: Houghton Mifflin, 1924.

Allport, F. H. A structuronomic conception of behavior: Individual and collective. I. Structural theory and the master problem of social psychology. Journal of Abnormal Social Psychology, 1962, 64, 3-30.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1971, 514-515.

Brennan, R. L. & Lockwood, R. E. A comparison of the Nedelsky and Angoff cutting score procedures using generalizablity theory. Applied Psychological Measurement, 1980, 4, 219-240.

Cartwright, D. & Harary, F. Structural balance: A generalization of Heider's theory. Psychological Review, 1956, 63, 277-293.

Festinger, L. A. A theory of social comparison processes. Human Relations, 1954, 7, 117-140.

Glass, G. V Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

Hambleton, R. K. Test score validity and standard-setting methods. In R. A. Berk (ed.), Criterion-Referenced Measurement: The State of the Art. Baltimore: Johns Hopkins University Press, 1980, 80-123.

Hambleton, R. K. & Eignor, D. R. Competency test development, validation, and standard setting. In R. M. Jaeger and C. K. Tittle (eds.), Minimum Competency Achievement Testing, Berkeley, CA: McCutchan, 1980, 367-396.

Hambleton, R. K., Swaminathan, H., Algina, J. & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Heider, F. Social perception and phenomenal causality. Psychological Review, 1944, 51, 358-374.

Heider, F. Attitudes and cognitive organization. Journal of Psychology, 1946, 21, 107-112.

Heider, F. The Psychology of Interpersonal Relations. New York: Wiley, 1958.

Jaeger, R. M. Measurement consequences of selected standard-

setting models. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April, 1976.

Jaeger, R. M. A proposal for setting a standard on the North Carolina high school competency test. Paper presented at the spring meeting of the North Carolina Association for Research in Education, Chapel Hill, 1978.

Jaeger, R. M. Measurement consequences of selected standard-setting models. In M. A. Bunda and J. R. Sanders (eds.), Practices and Problems in Competency-based Measurement, Washington, D. C.: National Council on Measurement in Education, 1979.

Jaeger, R. M. An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 1982, 4, 461-475.

Koffler, S. L. A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 1980, 17, 167-178.

Livingston, S. A. and Zieky, M. J. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, N. J.: Educational Testing Service, 1982.

Marquis, D. G. Individual responsibility and group decisions involving risk. Industrial Management Review, 1962, 3, 8-23.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 1976, 45, 133-158.

Myers, D. G., BAch, P. J. & Schreiber, F. B. Normative and informational effects of group interaction. Sociometry, 1974, 37, 275-286.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Schachter, S. The Psychology of Affiliation. Palo Alto, CA: Stanford University Press, 1959.

Shepard, L. Standard-setting issues and methods. Applied Psychological Measurement, 1980, 4, 447-467.

Sherif, M. Group influences upon the formation of norms and attitudes. In T. M. Newcomb & E. L. Hartley (eds.), Readings in Social Psychology. (1st. ed.) New York: Holt, 1947.

Stoner, J. A. F. A comparison of individual and group decisions involving risk. Unpublished master's thesis, School of Industrial Management, Massachusetts Institute of Technology, 1961.

Wallach, N. & Kogan, M. A. Sex differences and judgment processes. Journal of Personality, 1959, 27, 555-564.

Wallach, N. & Kogan, M. A. The roles of information, discussion, and consensus in group risk taking. Journal of Experimental Social Psychology, 1965, 1, 1-19.

Table 1. Means and Variances of Standards Recommended by the Discussion Group and the Silence Group on the Reading Subtest of the National Teacher Examinations.

Group	Type of Participant	First Judgment Session		Second Judgment Session	
		Mean	Variance	Mean	Variance
Discussion	College/Univ. (4)*	14.77	20.56	14.35	6.11
	Public School(3)	15.48	13.27	14.05	2.57
	Entire Group (7)	15.07	14.85	14.22	3.94
Silence	College/Univ.(4)	13.37	7.71	15.52	1.94
	Public School(3)	14.32	3.06	13.61	0.67
	Entire Group (7)	13.78	5.13	14.70	2.23
Both Groups	All Participants	14.42	9.67	14.46	2.91

*Numbers in parentheses denote sample sizes.

Table 2. Means and Variances of Standards Recommended by the Discussion Group and the Silence Group on the Social Studies Subtest of the National Teacher Examinations.

Group	Type of Participant	First Judgment Session		Second Judgment Session	
		Mean	Variance	Mean	Variance
Discussion	College/Univ. (4)*	8.98	8.85	9.84	5.66
	Public School(4)	17.25	11.24	13.62	0.19
	Entire Group (8)	13.12	28.16	11.73	6.60
Silence	College/Univ.(2)	16.64	1.63	16.30	0.13
	Public School(4)	14.24	32.42	15.22	16.56
	Entire Group (6)	15.04	21.32	15.58	10.27
Both Groups	All Participants	13.94	24.34	13.38	11.42

*Numbers in parentheses denote sample sizes.

Table 3. Repeated Measures Analyses of Variance of Recommended Standards on the Reading Subtest and Social Studies Subtest of the National Teacher Examinations.

Source	Reading Subtest d.f.	MS	F	Social Studies Subtest d.f.	MS	F
Between Subjects						
Type of Treatment	1	1.15	<1.00	1	57.15	2.94
Type of Participant	1	0.13	<1.00	1	56.89	2.92
Treatment by Type of Participant	1	0.81	<1.00	1	96.64	4.97**
Residual	10	11.90		10	19.45	
Within Subjects						
Session Number	1	0.01	<1.00	1	2.18	<1.00
Treatment by Session Number	1	5.56	1.91	1	6.36	1.99
Type of Participant by Session	1	6.41	2.20	1	7.77	2.43
Treatment by Type of Participant by Session Number	1	1.46	<1.00	1	13.53	4.24*
Residual	10	2.91		10	3.20	

* denotes $p<0.10$
** denotes $p<0.05$

Figure 1. Social Studies Interactions